Breakout Session 7: Track A

Making Parkinson's Disease Data AI-Ready for Cloud-Outsourced Machine Learning Research with Differential Privacy

Dr. Shigang Chen
Professor, University of Florida

Project Title: SCH: Enabling Data Outsourcing and Sharing for AI-powered Parkinson's Research

NOT-OD-22-067 Supplement Title: Making Parkinson’s Disease Data AI-Ready for Cloud-Outsourced Machine Learning Research with Differential Privacy

Presenter and PI: Shigang Chen, University of Florida

Summary of the Main Project

Outsource medical data and
AI training to cloud

Data privacy ?
Homomorphic encryption
Secure multiparty computation
Differential privacy
Federated model training
…

Data masking and masked model training

Research Aim 1: Experimental studies of building cloud-based masked ANN (artificial neural network) models for PD (Parkinson’s disease) quality-of-life prediction and parkinsonism classification

Research Aim 2: Theoretical studies on data privacy, inference accuracy, and training performance of cloud-based masked ANN models

Summary of the Supplement Project

Theoretical Extension: From matrix masking under an in-house privacy model to matrix masking plus noise addition under the well-accepted differential privacy model

Experimental Extension: From training Parkinson's diagnosis model based on matrix-masked data to training Parkinson's diagnosis model based on differentially private masked+noised data

Privacy-protected AI-ready Data: Transforming patient data with matrix masking and noise addition for outsourced deep learning in the cloud

Achieving Differential Privacy

Matrix Masking: Y = A X, achieving statistical data privacy [1, 2]

Noise Addition: Y = A + C, where $\mathbf{C} \sim NI_{n\times p}(0, \sigma^2)$, achieving differential privacy with noise level

$$\sigma \geq \frac{\bar{\gamma}\delta}{\varepsilon}(1 + \frac{1}{2\bar{\gamma}_\delta^2})$$

Matrix Masking + Noise Addition: Y = A (X + C) or Y = AX + C, where $\mathbf{C} \sim NI_{n\times p}(0, \sigma^2)$, achieving differential privacy with noise level

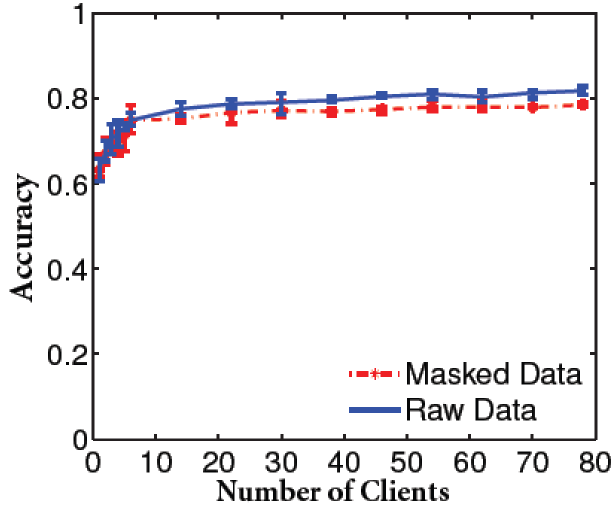
$$\sigma \geq \sqrt{\frac{2n - p + \ln(\frac{1}{\delta})}{2(n-p)}}\frac{3\sqrt[4]{p}}{\sqrt{\varepsilon}}.$$

Achieving Differential Privacy - continue

Table 1 Comparison of σ bounds (13) for setting (A) versus (14) for setting (B).

ε	δ	p	n	Setting (A) necessary (12)	Setting (A) sufficient (13)	Setting (B) sufficient (14)	Ratio of (13)/(14)
0.100	0.010	1	100	23.3	25.4	6.9	4
			10000	23.3	25.4	6.4	4
		5	100	23.3	25.4	9.5	3
			10000	23.3	25.4	8.9	3
		20	100	23.3	25.4	13.1	2
			10000	23.3	25.4	12.1	2
	0.001	1	100	30.9	32.5	7.1	5
			10000	30.9	32.5	6.4	5
		5	100	30.9	32.5	9.8	3
			10000	30.9	32.5	8.9	4
		20	100	30.9	32.5	13.5	2
			10000	30.9	32.5	12.1	3
0.010	0.010	1	100	232.6	254.1	21.8	12
			10000	232.6	254.1	20.2	13
		5	100	232.6	254.1	30.2	8
			10000	232.6	254.1	28.0	9
		20	100	232.6	254.1	41.5	6
			10000	232.6	254.1	38.3	7
	0.001	1	100	309.0	325.2	22.4	15
			10000	309.0	325.2	20.2	16
		5	100	309.0	325.2	31.0	10

Experimental Findings
- case of Y = AX



(b) POP data set.

Experimental findings [3, 4]

- Using masked data to train Parkinson’s disease models in cloud produces prediction accuracy close to models trained from raw patient data
- Method may be applicable to other diseases

Theoretical findings

- Difference between masked-data models and raw-data models are asymptotically bounded
- Data masking with random orthogonal transformation can achieve differential privacy with much smaller noise addition
- Data masking can work with federated learning
- Efficient data masking across multiple medical data sources is possible

Experimental Findings
- case of Y = A (X + C) or Y = A X + C

4908 patients and 11 attributes (after one-hot is 36 attributes)

sigma	0	0.1	0.3	0.5	1	1.5	2	2.5	3
Y=A(X+C)	76.56	76.03	72.73	68.92	56.65	43.44	40.93	40.89	41.1
Y=AX+C	76.84	75.79	73.28	69.69	55.48	40.05	40.05	40.05	40.09

12828 patients and 50 attributes (after one-hot is 118 attributes)

sigma	0	0.1	0.3	0.5	1	1.5	2	2.5	3
Y=A(X+C)	100	100	100	100	99.38	92.59	67.84	54.86	55.33
Y=AX+C	100	100	100	100	99.35	88.46	60.85	51.71	53.16

Findings

- Noise level significantly affects the accuracy of the model
- Increasing the size of the training data can compensate for the increase of noise (which means better differential privacy)
- Matrix masking and large data size make it feasible to outsource differentially private data for cloud-based deep learning

Privacy-protected AI-ready Data

We produced a privacy-protected AI-ready data set with matrix masking and noise addition.

We performed a case study to validate the feasibility of outsourcing privacy-protected data to the cloud for training diagnosis models with deep learning.

We investigated various approaches for filling in the missing data in order to increase the data size.

Challenges

Theoretical Challenge

- The challenge was that it was very difficult to derive the noise bound of $Y = A(X+C)$ for differential privacy and we could only derive an upper bound, which was not tight.
- We have been continuously working on this problem with a series of improving upper bounds. We suspect that the real bound is still much tighter.

Experimental and Data Preparation Challenge

- The challenge was that, in building cloud-based diagnosis models, larger noise was preferred for better privacy, yet the accuracy of the models deteriorated quickly with increasing noise.
- We addressed this issue in two ways: lowering the noise bound through theoretical work and increasing the number of usable patient records by exploring novel approaches of filling in the missing data.

Future Work

Theoretical work

- We will continuously work on deriving a tighter noise bound, such that we can reduce the noise level and improve the model accuracy, under the same differential privacy requirement.

Experimental work

- We will find novel methods to improve the model accuracy for cloud-based deep learning of Parkinson's disease diagnosis.
- We will try out other datasets based on our developed privacy-preserving data sourcing methods.

medical data outsourcing and sharing